

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BÙI ĐỨC VIỆT

**PHÂN CỤM DỮ LIỆU CHO NHẬN DẠNG ẢNH
SỬ DỤNG MẠNG NƠON**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN, NĂM 2012

LỜI CẢM ƠN

Trước tiên em gửi lời cảm ơn chân thành sâu sắc tới các thầy cô giáo ở Viện Công nghệ thông tin Việt Nam, các thầy cô trong trường Đại học Công nghệ thông tin & Truyền thông - Đại học Thái Nguyên đã tận tình truyền đạt, giảng dạy cho em những kiến thức, kinh nghiệm quý báu trong suốt thời gian qua.

Đặc biệt em xin gửi lời cảm ơn đến PGS.TS Lê Bá Dũng đã tận tình giúp đỡ, trực tiếp chỉ bảo em trong suốt thời gian làm luận văn. Trong thời gian làm việc với Thầy, em không những tiếp thu thêm nhiều kiến thức bổ ích mà còn học được tinh thần làm việc, thái độ nghiên cứu khoa học nghiêm túc, hiệu quả. Đây là những điều rất cần thiết cho em trong quá trình học tập và công tác.

Sau cùng xin gửi lời cảm ơn chân thành tới gia đình, bạn bè đã động viên, đóng góp ý kiến và giúp đỡ trong quá trình học tập, nghiên cứu và hoàn thành đề tài này.

Thái Nguyên, tháng 10 năm 2012

Học viên

Bùi Đức Việt

MỤC LỤC

MỤC LỤC.....	3
DANH MỤC CÁC KÝ HIỆU, CÁC TỪ VIẾT TẮT.....	6
DANH MỤC CÁC HÌNH VẼ.....	7
LỜI NÓI ĐẦU	9
CHƯƠNG 1. GIỚI THIỆU VỀ KHAI PHÁ DỮ LIỆU	11
1.1. Khái niệm khai phá dữ liệu	11
1.2. Kiến trúc của một hệ thống khai phá dữ liệu	11
1.3 Các giai đoạn của quá trình khai phá	13
1.4. Các phương pháp khai phá dữ liệu.....	14
1.5. Các cơ sở dữ liệu phục vụ cho khai phá dữ liệu.....	16
1.6. Các ứng dụng của khai phá dữ liệu	17
1.7. Các thách thức và khó khăn trong khai phá dữ liệu	17
1.8 Mạng nơron cho khai phá dữ liệu.....	18
CHƯƠNG 2. TỔNG QUAN VỀ PHÂN CỤM DỮ LIỆU	20
2.1. Khái niệm và mục tiêu của phân cụm dữ liệu	20
2.1.1. Khái niệm về phân cụm dữ liệu	20
2.1.1.1. Mục tiêu của phân cụm dữ liệu	20
2.1.1.2. Các yêu cầu đối với kỹ thuật phân cụm dữ liệu.....	21
2.1.1.3. Các kiểu dữ liệu và các thuộc tính trong phân cụm.....	23
2.2. Một số thuật toán trong phân cụm dữ liệu.....	25
2.2.1. Các thuật toán trong phân cụm phân hoạch.....	25
2.2.2. Các thuật toán trong phân cụm phân cấp.....	31
2.2.3. Các thuật toán phân cụm dựa trên mật độ	33
2.2.4. Phân cụm dựa trên lưới.....	34

2.2.5. Phân cụm dựa trên mô hình	35
2.2.6. Phân cụm có dữ liệu ràng buộc.....	36
2.3. Phân cụm cụm mờ	37
2.3.1. Tổng quan về phân cụm mờ	37
2.3.2. Các thuật toán phân cụm mờ	38
CHƯƠNG 3: ỨNG DỤNG MẠNG NƠON KOHONEN CHO PHÂN CỤM DỮ	
LIỆU.....	42
3.1. Giới thiệu chung về mạng nơon.....	42
3.1.2. Mô hình Nơon sinh học	42
3.1.3. Mô hình Nơon nhân tạo.....	44
3.1.4. Mô hình Mạng Nơon nhân tạo	46
3.1.5. Đặc trưng của Mạng Nơon	50
3.1.6. Phân loại mạng	51
3.2.3. Thuật toán của mạng SOM	59
3.2.4. Một vài biến thể của giải thuật SOM.....	65
3.2.5. Một số ứng dụng của SOM.....	66
CHƯƠNG 4: CÀI ĐẶT CHƯƠNG TRÌNH THỬ NGHIỆM	67
4.1 Giới thiệu.....	67
4.2 Mạng Nơon SOM cho phân cụm ảnh	68
Thiết kế mạng	68
Thuật toán học mạng	68
4.2 Giới thiệu môi trường cài đặt	70
4.3 Giới thiệu giao diện chương trình	70
4.3.1 Thử nghiệm 1	70
4.3.2 Thử nghiệm 2.....	73
4.4 Hạn chế của giải thuật SOM khi áp dụng phân cụm màu trên ảnh.....	74
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	77

TÀI LIỆU THAM KHẢO.....	77
-------------------------	----

DANH MỤC CÁC KÝ HIỆU, CÁC TỪ VIẾT TẮT

CSDL	Cơ sở dữ liệu
PCDL	Phân cụm dữ liệu
KPDL	Khai phá dữ liệu
BNU	Phần tử noron chiến thắng
MLP	MultiLayer Perception
BAM	Bidirectional Associative Memory
SOM	Self Organizing Map
VQ	Vector Quantization
LVQ	Learning Vector Quantization
MST	Minimal Spanning Tree

DANH MỤC CÁC HÌNH VẼ

	Trang
Hình 1.1: Kiến trúc một hệ thống khai phá dữ liệu	11
Hình 1.2: Quá trình khai phá dữ liệu	12
Hình 2.1 : Biểu đồ các dạng dữ liệu	22
Hình 2.2: biểu đồ quy mô dữ liệu	22
Hình 2.3: Cấu trúc phân cấp	27
Hình 2.4: Các cách mà cụm có thể đưa ra	28
Hình 2.5: Thiết lập để xác định danh giới các cụm ban đầu	30
Hình 2.6: Tính toán trọng tâm các cụm mới	31
Hình 2.7: Khái quát thuật toán Cure	36
Hình 2.8: Các cụm dữ liệu được khám phá bởi thuật toán Cure	37
Hình 2.9: Hình dạng các cụm được tạo bởi thuật toán DBSCAN	38
Hình 3.1: Mô hình nơron sinh học	49
Hình 3.2: Mô hình nơron nhân tạo cơ bản	53
Hình 3.2: Mô hình mạng nơron 3 lớp	52
Hình 3.3: Mô hình học giám sát	55
Hình 3.4: Mô hình học không giám sát	55
Hình 3.5: Mô hình mạng perceptron một lớp	58
Hình 3.6: Mô hình Mạng perceptron nhiều lớp	58
Hình 3.7: Mô hình mạng hồi quy một lớp	59
Hình 3.8: Cấu trúc của mạng Hopfield	60
Hình 3.9: Cấu trúc của mạng BAM	60
Hình 3.10: Mô hình Mạng Nơron Kohonen	63
Hình 3.11: Mô hình Mạng Nơron Kohonen thông thường	65
Hình 3.12: Phần tử nơron chiến thắng BMU	66

Hình 3.13: Các vùng lân cận	67
Hình 4.1: Giao diện chương trình	69
Hình 4.2: Khởi tạo mạng ngẫu nhiên	70
Hình 4.3: Xác định BMU	70
Hình 4.4: Kết quả gom cụm	71
Hình 4.5: Giao diện chọn ảnh để phân cụm	71
Hình 4.6: Kết quả sau khi phân cụm	72

LỜI NÓI ĐẦU

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin trong nhiều lĩnh vực của đời sống, kinh tế xã hội nhiều năm qua cũng đồng nghĩa với lượng dữ liệu đã được các cơ quan thu thập và lưu trữ ngày một tích lũy nhiều lên. Nếu cho rằng, điện tử và truyền thông chính là bản chất của khoa học điện tử thì dữ liệu, thông tin, tri thức là tiêu điểm của một lĩnh vực mới để nghiên cứu và ứng dụng, đó là khám phá tri thức và khai phá dữ liệu.

Thông thường, chúng ta coi dữ liệu là một chuỗi các bits, hoặc các số và các ký hiệu hay các đối tượng với một ý nghĩa nào đó khi gửi cho một chương trình dưới một dạng nhất định. Các bit thường được sử dụng để đo thông tin, và xem nó như là dữ liệu đã loại bỏ phần tử dư thừa, lặp lại và rút gọn tới mức tối thiểu để đặc trưng một cách cơ bản cho dữ liệu. Tri thức được xem như là những thông tin tích hợp, bao gồm các sự kiện và các mối quan hệ giữa chúng đã được nhận thức, khám phá hoặc nghiên cứu. Nói cách khác, tri thức có thể được coi là dữ liệu ở mức độ cao của sự trừu tượng và tổng quát.

Khám phá tri thức hay phát hiện tri thức trong CSDL là quy trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: phân tích, tổng hợp, hợp thức, khả ích và có thể hiểu được.

Khai phá dữ liệu là một bước trong quá trình khám phá tri thức, gồm các thuật toán khai thác dữ liệu chuyên dùng dưới một số quy định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói cách khác, mục tiêu của khai phá dữ liệu là tìm kiếm các mẫu hoặc các mô hình tồn tại trong CSDL nhưng ẩn trong khối lượng lớn dữ liệu.

Phân cụm dữ liệu (PCDL) là quá trình nhóm một tập các đối tượng tương tự nhau trong tập dữ liệu vào các cụm sao cho các đối tượng thuộc cùng một cụm là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng. Phân cụm dữ

liệu là một ví dụ của phương pháp học không có thầy. Không giống như phân lớp dữ liệu, phân cụm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân cụm dữ liệu là một cách học bằng quan sát, trong khi phân lớp dữ liệu là học bằng ví dụ...

Hiện nay, các phương pháp phân cụm trên đã và đang được phát triển và áp dụng nhiều trong các lĩnh vực khác nhau và đã có một số nhánh nghiên cứu được phát triển trên cơ sở của các phương pháp đó như:

Phân cụm thống kê: Dựa trên các khái niệm phân tích hệ thống, nhánh nghiên cứu này sử dụng các độ đo tương tự để phân hoạch các đối tượng, nhưng chúng chỉ áp dụng cho các dữ liệu có thuộc tính số.

Phân cụm khái niệm: Kỹ thuật này được phát triển áp dụng cho dữ liệu hạng mục, chúng phân cụm các đối tượng theo các khái niệm mà chúng xử lí.

Phân cụm mờ: Sử dụng kỹ thuật mờ để PCDL. Các thuật toán thuộc loại này chỉ ra lược đồ phân cụm thích hợp với tất cả các hoạt động đời sống hàng ngày, chúng chỉ xử các dữ liệu không chắc chắn.

Luận văn gồm có 4 chương:

Chương 1: Giới thiệu về khai phá dữ liệu

Chương 2: Tổng quan về phân cụm dữ liệu

Chương 3: Ứng dụng mạng Nơron Kohonen cho phân cụm dữ liệu

Chương 4: cài đặt thử nghiệm

Luận văn đã trình bày một số vấn đề về phân cụm - một trong những kỹ thuật cơ bản để khai phá dữ liệu và ứng dụng phân cụm cho nhận dạng ảnh sử dụng mạng nơron. Đây là hướng nghiên cứu có triển vọng chỉ ra những sơ lược trong việc hiểu và khai thác CSDL khổng lồ, khám phá thông tin hữu ích ẩn trong dữ liệu; hiểu được ý nghĩa thực tế của dữ liệu và ứng dụng vào bài toán cụ thể.